Theory of Language Modeling Part I: A Markov Categorical Framework for Language Modeling

Yifan Zhang
yif-zhang@outlook.com

March 29, 2025

Abstract

Auto-regressive (AR) language models underpin state-of-the-art natural language processing, factorizing sequence probabilities as $P_\theta(\mathbf{w}) = \prod_t P_\theta(w_t|\mathbf{w}_{<t})$. While empirically powerful, their internal information processing pathways, particularly the mapping from context $\mathbf{w}_{<t}$ to the next-token predictive distribution $P_\theta(\cdot|\mathbf{w}_{<t})$, lack a comprehensive theoretical characterization. This work introduces a rigorous analytical framework grounded in the theory of Markov Categories (MCs), specifically employing the category **Stoch** whose objects are standard Borel spaces (e.g., sequence spaces $\mathcal{V}^*$, representation spaces $\mathcal{H} \cong \mathbb{R}^{d_{\text{model}}}$, finite vocabularies $\mathcal{V}$) and morphisms are Markov kernels (stochastic maps). We formally model the AR generation step as a composite Markov kernel $k_{\text{gen},\theta} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$, resulting from the composition $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$. Here k_{emb} and k_{bb} represent the typically deterministic context embedding and backbone transformations yielding the final hidden state $h_t \in \mathcal{H}$, while $k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ is the stochastic kernel mapping h_t to $P_\theta(\cdot|\mathbf{w}_{<t})$. A key contribution is leveraging the enrichment of **Stoch** with statistical divergences D (e.g., D_{KL}, d_{TV}, f-divergences) which satisfy the Data Processing Inequality (DPI). Building upon Perrone's work [15], we utilize the intrinsically defined categorical information measures of kernel entropy $\mathcal{H}_D$ and mutual information I_D. This enables the definition of principled, quantitative metrics to dissect the AR process: (1) **Representation Fidelity**: Measuring the distinguishability of context-property-conditioned hidden state distributions $p_{H_t|s_1}$ and $p_{H_t|s_2}$ via the divergence $D_{\mathcal{H}}(p_{H_t|s_1} \| p_{H_t|s_2})$. (2) **Categorical Mutual Information**: Quantifying statistical dependencies via I_D, specifically state-prediction relevance $I_D(H_t; W_t)$ and temporal state coherence $I_D(H_t; H_{t+1})$, using joint states constructed categorically. (3) **Kernel Stochasticity**: Assessing the intrinsic randomness of the final prediction stage via the categorical entropy $\mathcal{H}_D(k_{\text{head}})$. (4) **Information Flow Bounds**: Applying the DPI inherent to I_D within the Markov chain $S \to H_t \to W_t$ to establish fundamental limits, e.g., $I_D(S; H_t) \geq I_D(S; W_t)$. This framework provides a unified, mathematically rigorous, and compositional perspective, distinct from layer-specific probes or purely empirical methods, offering tools from category theory, probability, and information geometry to analyze information compression, representation structure, and the statistical underpinnings of large language models.

1 Introduction

Auto-regressive language models (AR LMs), particularly those based on the Transformer architecture [20, 17, 4], have achieved remarkable success, defining the state-of-the-art in natural language generation and demonstrating impressive few-shot learning capabilities. These models operate by sequentially predicting the next token in a sequence based on the preceding context. Formally, given a sequence $\mathbf{w} = w_1 \dots w_L$ with tokens w_i from a finite vocabulary $\mathcal{V}$, the model learns a parameterized probability distribution P_θ that factorizes as:

$$P_\theta(\mathbf{w}) = \prod_{t=1}^{L} P_\theta(w_t|\mathbf{w}_{<t}) \tag{1}$$

where $\mathbf{w}_{<t} \coloneqq w_1 \dots w_{t-1}$ is the context sequence, and θ denotes the model parameters, typically optimized by minimizing the negative log-likelihood (cross-entropy) on vast text corpora. The core computational step is the mapping from a context $\mathbf{w}_{<t}$ to the conditional probability distribution $P_\theta(\cdot|\mathbf{w}_{<t})$ over $\mathcal{V}$ for the next token w_t.

Despite their empirical triumphs, a deep theoretical understanding of the internal information processing pathways and representation learning mechanisms within these large-scale models remains largely incomplete [12, 9, 6]. Current analytical approaches often rely on empirical probes of neural activations [8], correlation studies with linguistic features, or detailed analyses of specific architectural components like attention heads [14]. While insightful, these methods often lack a unified, mathematically principled framework to capture the compositional nature of the computation and the inherent stochasticity involved in the generation process.

This paper introduces a rigorous analytical framework for the AR generation step $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$ rooted in the theory of Markov Categories (MCs) [5, 7]. MCs provide an abstract algebraic setting tailored for reasoning about systems involving probability, causality, conditioning, and information flow using the language of category theory. We specifically leverage the category **Stoch**, a canonical MC whose objects are standard Borel spaces (a well-behaved class of measurable spaces suitable for probability theory, encompassing finite sets like $\mathcal{V}$, Euclidean spaces like $\mathbb{R}^d$, and sequence spaces like $\mathcal{V}^*$) and whose morphisms are Markov kernels, representing conditional probability distributions [10, 7].

The strength of the MC framework lies in its inherent compositionality, mirroring the layered structure of deep neural networks, its native handling of probability and stochastic transformations, and its capacity for defining fundamental information-theoretic quantities in a principled manner. We model the complex computation within an AR LM as a sequence of morphisms (Markov kernels) composed in **Stoch**. Specifically, the generation process is factored as:

$$k_{\text{gen},\theta} \coloneqq k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{V}, \mathcal{P}(\mathcal{V})) \tag{2}$$

Here:

- $(\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$ is the standard Borel space of context sequences.
- $k_{\text{emb}} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}))$ is typically a deterministic kernel (corresponding to the token embedding and initial positional encoding function) mapping to an intermediate space $(\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}))$ of sequence representations (e.g., sequences of vectors).

- $k_{\text{bb}} : (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}})) \to (\mathcal{H}, \mathscr{B}(\mathcal{H}))$ represents the deterministic transformation enacted by the model's backbone (e.g., Transformer layers, potentially incorporating mechanisms like RoPE [18]) mapping to the final hidden state space $(\mathcal{H}, \mathscr{B}(\mathcal{H}))$, where $\mathcal{H} \cong \mathbb{R}^{d_{\text{model}}}$.
- $k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ is the generally stochastic kernel corresponding to the LM head (linear layer plus softmax), mapping the final hidden state $h_t \in \mathcal{H}$ to the predictive distribution over the vocabulary space $(\mathcal{V}, \mathcal{P}(\mathcal{V})) = (\mathcal{V}, \mathcal{P}(\mathcal{V}))$.

The composite kernel $k_{\text{gen},\theta}$ precisely captures the mapping $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$.

A crucial aspect of our framework is the enrichment of **Stoch** with a statistical divergence D (e.g., D_{KL}, d_{TV}, Rényi α-divergence) [2, 15, 16]. Divergences quantify the dissimilarity between probability distributions and, critically, satisfy the Data Processing Inequality (DPI): processing through any Markov kernel k cannot increase the divergence between input distributions. Building on this, Perrone [15] introduced intrinsic, categorical definitions of entropy $\mathcal{H}_D$ and mutual information I_D associated with a divergence D. These definitions inherit fundamental properties like the DPI by construction.

Leveraging this divergence-enriched Markov Category $(\mathbf{Stoch}, D)$, this paper makes the following contributions:

1. **Formal MC Model:** We provide a precise formulation of the AR LM's single-step generation process as a composite Markov kernel (Equation (2)) in **Stoch**.
2. **Categorical Information Metrics:** We propose and define novel metrics to analyze information processing, based directly on categorical entropy and mutual information:
   - **Representation Divergence** (RepDiv_D): Quantifies how well the hidden state h_t distinguishes context properties s by measuring $D(p_{H_t|s_1} \| p_{H_t|s_2})$.
   - **Categorical Mutual Information**: Measures statistical dependencies like state-prediction relevance $I_D(H_t; W_t)$ and temporal state coherence $I_D(H_t; H_{t+1})$.
   - **LM Head Categorical Entropy** ($\mathcal{H}_D(k_{\text{head}})$): Measures the intrinsic stochasticity of the final prediction kernel.
3. **Information Flow Bounds:** We utilize the inherent DPI satisfied by I_D within the Markov chain structure (e.g., Context Property $S \to H_t \to W_t$) to establish fundamental bounds, such as $I_D(S; H_t) \geq I_D(S; W_t)$, quantifying information preservation and loss.
4. **Conceptual Links:** We connect the MC framework to information geometry (Section 5) and propose interpretations of representation learning based on implicit spectral structuring (Section 6) and compression (Section 4).

This framework offers a principled, mathematically grounded, and compositional alternative to heuristic or purely empirical analysis methods. By providing tools rooted in category theory, probability, and information geometry, it aims to facilitate a deeper quantitative understanding of information compression, representation learning, and the statistical structure underpinning the success of modern AR language models. Such foundational tools are crucial for advancing the theoretical understanding of deep generative models, a key goal for the machine learning community.

The paper is organized as follows: Section 2 provides the necessary mathematical background on AR LMs, the category **Stoch**, divergences, and categorical information measures. Section 3 formally defines our proposed metrics within the MC framework. Sections 4 to 6 discuss interpretations related to compression, information geometry, and implicit structure learning. Section 7 discusses limitations and future work, followed by the conclusion in Section 8.

2 Background

This section reviews the essential concepts required for our framework: the representation of AR LM components as Markov kernels, the definition of the Markov category **Stoch**, and the enrichment of **Stoch** with statistical divergences leading to categorical definitions of entropy and mutual information.

2.1 Auto-Regressive Language Models as Composed Kernels

We model the single-step generation mapping $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$ as a composition of Markov kernels within the category **Stoch**. Standard Borel spaces are chosen as objects because they form a well-behaved class of measurable spaces (isomorphic to Borel subsets of Polish spaces) closed under countable products, sums, and containing standard examples like $\mathbb{R}^d$ and finite sets, ensuring measure-theoretic regularity [10].

The relevant measurable spaces are:

- Input context space: $(\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$, where $\mathcal{V}^*$ is the set of finite sequences over the vocabulary $\mathcal{V}$, equipped with a suitable σ-algebra making it standard Borel (e.g., considering it as a disjoint union of finite products $\mathcal{V}^n$).
- Initial sequence representation space: $(\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}))$, the space of initial vector sequences (e.g., $\bigcup_n (\mathbb{R}^{d_{\text{model}}})^n$), also equipped with a standard Borel structure.
- Final hidden state space: $(\mathcal{H}, \mathscr{B}(\mathcal{H}))$, typically $(\mathbb{R}^{d_{\text{model}}}, \mathscr{B}(\mathbb{R}^{d_{\text{model}}}))$.
- Output vocabulary space: $(\mathcal{V}, \mathcal{P}(\mathcal{V})) = (\mathcal{V}, \mathcal{P}(\mathcal{V}))$, a finite measurable space.

The process decomposes into three kernels:

1. **Embedding Layer Kernel ($k_{\text{emb}} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}))$)**: This kernel encapsulates the initial processing of the discrete input sequence $\mathbf{w}_{<t} \in \mathcal{V}^*$. It typically involves applying a token embedding function $\mathcal{E} : \mathcal{V} \to \mathbb{R}^{d_{\text{model}}}$ to each token w_i and potentially incorporating absolute positional encodings. Let $f_{\text{emb}} : \mathcal{V}^* \to \mathcal{H}_{\text{seq\_emb}}$ denote the overall deterministic function computing the initial sequence representation $E_{<t}$. Since this mapping is deterministic, the kernel k_{emb} is defined via the Dirac measure $\delta_\cdot$:

$$k_{\text{emb}}(\mathbf{w}_{<t}, A) := \delta_{f_{\text{emb}}(\mathbf{w}_{<t})}(A) = \mathbf{1}_A(f_{\text{emb}}(\mathbf{w}_{<t})), \quad \text{for } A \in \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}). \tag{3}$$

This is a morphism in **Stoch**.

2. **Backbone Transformation Kernel ($k_{\text{bb}} : (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}})) \to (\mathcal{H}, \mathscr{B}(\mathcal{H}))$)**: This kernel represents the core computation, usually a deep neural network like a Transformer stack. Let $f_{\text{bb}} : \mathcal{H}_{\text{seq\_emb}} \to \mathcal{H}$ be the function mapping the initial sequence representation $E_{<t}$ to the final hidden

state $h_t \in \mathcal{H}$ (often the output vector at the last sequence position). This function incorporates complex operations like multi-head self-attention and feed-forward layers. Relative positional information, such as Rotary Position Embeddings (RoPE) [18], is implemented *within* the function f_{bb} by modifying attention computations based on token positions. Assuming the backbone computation is deterministic for a given $E_{<t}$ and parameters θ, the kernel k_{bb} is also deterministic:

$$k_{\text{bb}}(E_{<t}, B) := \delta_{f_{\text{bb}}(E_{<t})}(B) = \mathbf{1}_B(f_{\text{bb}}(E_{<t})), \quad \text{for } B \in \mathscr{B}(\mathcal{H}). \tag{4}$$

This is also a morphism in **Stoch**.

3. **LM Head Kernel ($k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$)**: This final kernel maps the summary hidden state $h_t \in \mathcal{H}$ to a probability distribution over the finite vocabulary $\mathcal{V}$. Typically, h_t is passed through a linear layer ($f_{\text{head}} : \mathcal{H} \to \mathbb{R}^{|\mathcal{V}|}$) producing logits $\mathbf{z} = f_{\text{head}}(h_t)$, followed by the $\mathrm{softmax}$ function: $P(w|h_t) = [\mathrm{softmax}(\mathbf{z})]_w$. This defines a genuinely stochastic Markov kernel:

$$k_{\text{head}}(h, A) := \sum_{w \in A} [\mathrm{softmax}(f_{\text{head}}(h))]_w \quad \text{for } h \in \mathcal{H}, A \subseteq \mathcal{V}. \tag{5}$$

This kernel maps each point h in the representation space to a probability measure on the discrete space $\mathcal{V}$, satisfying the required measurability conditions. It is a morphism in **Stoch**.

The overall generation kernel $k_{\text{gen},\theta} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ is the composition $k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ in **Stoch**, representing the model's learned conditional probability map $P_\theta(\cdot|\mathbf{w}_{<t})$.

2.2 Markov Categories and Stoch

Markov Categories provide an axiomatic framework for probability and stochastic processes using category theory [7].

Definition 2.1 (Markov Category [7])**.** *A Markov category $(\mathcal{C}, \otimes, I)$ is a symmetric monoidal category where each object X is equipped with a commutative comonoid structure $(\Delta_X : X \to X \otimes X, !_X : X \to I)$ that is natural in X, and the monoidal unit I is a terminal object (the causality axiom: $!_X$ is the unique map $X \to I$).*

Morphisms $k : X \to Y$ are interpreted as stochastic processes or channels transforming systems of type X to type Y. Composition $h \circ k$ denotes sequential processing, $k \otimes h$ parallel processing. The comonoid maps Δ_X (copy) and $!_X$ (discard) model the duplication and deletion of information. States (probability distributions) on X are morphisms $p : I \to X$.

The key example for our purposes is the category **Stoch**.

Definition 2.2 (Category **Stoch** [7, 15])**.** *The Markov category* **Stoch** *is defined by:*

- ***Objects**: Standard Borel spaces $(X, \mathscr{B}(X))$. The monoidal unit I is a singleton space $(\{\star\}, \{\emptyset, \{\star\}\})$.*
- ***Morphisms**: Markov kernels $k : X \to Y$. A map $k : X \times \mathscr{B}(Y) \to [0, 1]$ where $k(x, \cdot)$ is a probability measure on Y for each $x \in X$, and $k(\cdot, A)$ is a measurable function on X for each $A \in \mathscr{B}(Y)$.*
- ***Composition**: Given $k : X \to Y$ and $h : Y \to Z$, the composite $h \circ k : X \to Z$ is $(h \circ k)(x, C) := \int_Y h(y, C)\, k(x, \mathrm{d}y)$ (Chapman-Kolmogorov). Identity $\mathrm{id}_X(x, A) = \delta_x(A)$.*

- ***Monoidal Product*** **($\otimes$)**: *Product space $(X \times Y, \mathscr{B}(X) \otimes \mathscr{B}(Y))$ with the product σ-algebra. Product kernel $(k \otimes h)((x,y),\cdot) \coloneqq k(x,\cdot) \otimes h(y,\cdot)$ (product measure).*
- ***Symmetry***: *Swap map $\sigma_{X,Y} : X \otimes Y \to Y \otimes X$ is $\sigma_{X,Y}((x,y),\cdot) = \delta_{(y,x)}$.*
- ***Comonoid Structure***: *Copy $\Delta_X : X \to X \otimes X$ is $\Delta_X(x,\cdot) = \delta_{(x,x)}$. Discard $!_X : X \to I$ maps to the unique point measure on I, $!_X(x,\{\star\}) = 1$.*
- ***Causality***: *I is terminal, $!_Y \circ k = !_X$ holds, reflecting probability normalization.*

Remark 2.3 (Interpretation)**.** *In* **Stoch***, objects represent the types of random outcomes (e.g., sequences, vectors, tokens). Morphisms represent stochastic processes or channels mapping inputs to probability distributions over outputs. Deterministic functions $f : X \to Y$ correspond to deterministic kernels $k_f(x,\cdot) = \delta_{f(x)}$. States $p : I \to X$ correspond bijectively to probability measures $\mu_p \in \mathcal{P}(X)$ via $\mu_p(A) = p(\star, A)$. Marginalization arises from discarding information, e.g., for a joint state $p : I \to X \otimes Y$, the X-marginal is $p_X = (\mathrm{id}_X \otimes !_Y) \circ p$.*

2.3 Divergence Enrichment and Categorical Information Measures

The structure of **Stoch** is particularly powerful when enriched with a statistical divergence D, quantifying the dissimilarity between probability measures (states) $p, q : I \to X$, written $D_X(p\|q)$ [15]. Examples include KL divergence (D_{KL}), Total Variation (d_{TV}), Rényi divergences (D_α), and the broad class of f-divergences (D_f) [1, 13].

A fundamental property linking divergences and Markov kernels is the Data Processing Inequality (DPI), which holds for most standard divergences (e.g., f-divergences, Rényi $\alpha \in [0,\infty]$).

Theorem 2.4 (Data Processing Inequality (DPI))**.** *Let D be a statistical divergence satisfying the DPI. For any Markov kernel $k : X \to Y$ in* **Stoch** *and any pair of states $p, q : I \to X$:*

$$D_Y(k \circ p \| k \circ q) \leq D_X(p\|q) \tag{6}$$

Processing through k cannot increase the D-divergence between the distributions.

Based on this, Perrone [15] introduced categorical definitions of entropy and mutual information intrinsically tied to the divergence D and the MC structure.

Definition 2.5 (Categorical Entropy and Mutual Information [15])**.** *Let $(\mathbf{Stoch}, D)$ be enriched with a DPI-satisfying divergence D.*

1. *The* ***Categorical Entropy*** *of a kernel $k : X \to Y$ measures its intrinsic stochasticity:*

$$\mathcal{H}_D(k) \coloneqq D_{Y\otimes Y}\left(\Delta_Y \circ k \quad \| \quad (k \otimes k) \circ \Delta_X\right) \tag{7}$$

It compares two processes producing pairs in $Y \otimes Y$. The first ($\Delta_Y \circ k$) applies k once ($x \mapsto y \sim k(x,\cdot)$) and deterministically copies the output (y,y). The second ($(k \otimes k) \circ \Delta_X$) deterministically copies the input (x,x) and applies k independently to each component (y_1, y_2) where $y_1, y_2 \sim k(x,\cdot)$ are i.i.d. The divergence measures how different these two resulting joint distributions are, quantifying how far k is from being deterministic. If k is deterministic, $k = k_f$, both sides yield the same state (corresponding to $\delta_{(f(x),f(x))}$) and $\mathcal{H}_D(k_f) = 0$.

2. *The **Categorical Mutual Information** of a joint state $p : I \to X \otimes Y$ measures the statistical dependence between X and Y:*

$$I_D(p) \coloneqq D_{X\otimes Y}\left(p \quad \| \quad p_X \otimes p_Y\right) \tag{8}$$

where $p_X = (\mathrm{id}_X \otimes !_Y) \circ p$ and $p_Y = (!_X \otimes \mathrm{id}_Y) \circ p$ are the marginal states. $I_D(p)$ measures how far the joint state p is from the product of its marginals (representing independence), according to the geometry induced by D.

Remark 2.6 (Properties and Connections)**.** *When $D = D_{\mathrm{KL}}$, $I_{D_{\mathrm{KL}}}(p)$ recovers the standard Shannon mutual information $I(X;Y)$ for the joint distribution p. $\mathcal{H}_{D_{\mathrm{KL}}}(k)$ provides an intrinsic measure of the kernel's stochasticity, related to but distinct from average conditional Shannon entropy [15]. Crucially, these categorical definitions automatically satisfy the DPI. For instance, consider a state $p_{XY} : I \to X \otimes Y$ and a kernel $h : Y \to Z$. Let p_{XZ} be the state obtained by applying $\mathrm{id}_X \otimes h$ to p_{XY}. The DPI for D applied to the states involved in the definition of I_D implies $I_D(p_{XY}) \geq I_D(p_{XZ})$ [15, Prop. 4.8]. This reflects the principle that processing ($Y \to Z$) cannot increase information about X. Furthermore, information geometry [1] arises naturally: the Fisher-Rao metric is induced by the local quadratic approximation of the KL divergence, linking the divergence D to the underlying geometric structure of the space of probability measures.*

3 Markov Categorical Metrics

We now apply the Markov category framework $(\mathbf{Stoch}, D)$ to analyze the AR generation kernel $k_{\mathrm{gen},\theta} = k_{\mathrm{head}} \circ k_{\mathrm{bb}} \circ k_{\mathrm{emb}}$ (Equation (2)). We select a suitable statistical divergence D satisfying the Data Processing Inequality (DPI) (e.g., D_{KL}, d_{TV}, or more generally an f-divergence [1, 13]) and utilize the corresponding categorical information measures $\mathcal{H}_D$ and I_D (Equations (7) and (8)) to probe the information flow and transformations within the generation step. A particular focus is placed on the final hidden state $H_t \in \mathcal{H}$ and the stochastic prediction kernel k_{head}.

We operate within the probabilistic setting induced by a distribution over input contexts. Let P_{ctx} be a probability measure on the context space $(\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) = (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$. This corresponds to an initial *state* in the Markov category **Stoch**, represented by a morphism $p_{W_{<t}} : I \to (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$, where I is the monoidal unit (a singleton measurable space) and $p_{W_{<t}}(\star, A) = P_{\mathrm{ctx}}(A)$ for any $A \in \mathscr{B}(\mathcal{V}^*)$. Processing this initial state through the sequence of deterministic kernels k_{emb} and k_{bb}, and the stochastic kernel k_{head}, induces distributions (states) at subsequent stages:

- **Initial Sequence Embedding State**: Given $p_{W_{<t}} : I \to (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$, the distribution of the initial vector sequence representation $E_{<t} \in \mathcal{H}_{\mathrm{seq\_emb}}$ is given by the state $p_{E_{<t}} : I \to (\mathcal{H}_{\mathrm{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\mathrm{seq\_emb}}))$, defined as:

$$p_{E_{<t}} \coloneqq k_{\mathrm{emb}} \circ p_{W_{<t}}. \tag{9}$$

  Since k_{emb} corresponds to the deterministic function f_{emb}, the measure associated with $p_{E_{<t}}$ is the pushforward measure $(P_{\mathrm{ctx}}) \circ f_{\mathrm{emb}}^{-1}$.

- **Final Hidden State**: The distribution of the final hidden state $H_t \in \mathcal{H}$ is given by the state $p_{H_t} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H}))$:

$$p_{H_t} \coloneqq k_{\mathrm{bb}} \circ p_{E_{<t}} = (k_{\mathrm{bb}} \circ k_{\mathrm{emb}}) \circ p_{W_{<t}}. \tag{10}$$

As k_{bb} is also deterministic (representing f_{bb}), p_{H_t} corresponds to the pushforward measure $(P_{\text{ctx}}) \circ (f_{\text{bb}} \circ f_{\text{emb}})^{-1}$.

- **Predicted Next Token State**: The marginal distribution of the predicted next token $W_t \in \mathcal{V}$, averaged over all contexts according to P_{ctx}, is given by the state $p_{W_t} : I \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$:

$$p_{W_t} \coloneqq k_{\text{head}} \circ p_{H_t} = (k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}) \circ p_{W_{<t}} = k_{\text{gen},\theta} \circ p_{W_{<t}}. \tag{11}$$

Let μ_{H_t} be the measure on $\mathcal{H}$ associated with p_{H_t}. Then the measure associated with p_{W_t} on $\mathcal{V}$ is given by $\mu_{W_t}(A) = \int_{\mathcal{H}} k_{\text{head}}(h, A)\, \mu_{H_t}(\mathrm{d}h)$ for $A \subseteq \mathcal{V}$.

Using these rigorously defined states and the categorical information measures, we propose the following metrics.

3.1 Metric 1: Representation Divergence (Context Encoding Fidelity)

Goal: To quantify how effectively the distribution of the final hidden state H_t distinguishes between different underlying properties S of the input context $\mathbf{w}_{<t}$.

Setup: Consider a random variable S, defined on the probability space underlying P_{ctx}, representing a specific property of the context $\mathbf{w}_{<t}$ (e.g., topic membership $s \in \{s_1, s_2, \dots\}$, presence/absence of a feature). Assume we can condition the context distribution P_{ctx} on the value of S. Let $P_{\text{ctx}}(\cdot|S = s)$ denote the conditional probability measure on $(\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$. This corresponds to a conditional input state $p_{W_{<t}|s} : I \to (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$ for each value s of S. The conditional distribution of the hidden state H_t given $S = s$ is then represented by the state:

$$p_{H_t|s} \coloneqq (k_{\text{bb}} \circ k_{\text{emb}}) \circ p_{W_{<t}|s} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H})). \tag{12}$$

Let $\mu_{H_t|s}$ be the probability measure on $\mathcal{H}$ associated with the state $p_{H_t|s}$.

Metric: The Representation Divergence between contexts exhibiting properties s_1 and s_2 is defined as the statistical divergence D between the corresponding conditional hidden state measures:

$$\text{RepDiv}_D(s_1 \| s_2) \coloneqq D_{\mathcal{H}}(\mu_{H_t|s_1} \| \mu_{H_t|s_2}) \equiv D_{\mathcal{H}}(p_{H_t|s_1} \| p_{H_t|s_2}). \tag{13}$$

Here, $D_{\mathcal{H}}$ denotes the application of the divergence functional D to probability measures (states) on the measurable space $(\mathcal{H}, \mathscr{B}(\mathcal{H}))$.

Interpretation: A large value of $\text{RepDiv}_D(s_1 \| s_2)$ indicates that the measures $\mu_{H_t|s_1}$ and $\mu_{H_t|s_2}$ are highly distinguishable according to the chosen divergence D. This implies that the transformation $(k_{\text{bb}} \circ k_{\text{emb}})$ maps contexts with properties s_1 and s_2 to significantly different distributions in the representation space $\mathcal{H}$. The hidden state H_t thus serves as an effective statistical signature for distinguishing between properties s_1 and s_2. Conversely, a small divergence suggests that the representations generated from contexts with properties s_1 and s_2 are statistically similar, implying that the model either does not encode this specific distinction strongly in H_t or represents them in overlapping regions of $\mathcal{H}$. The choice of D influences the notion of distinguishability (e.g., D_{KL} emphasizes differences in likelihood ratios, while d_{TV} focuses on the maximal difference in probability assigned to any event).

Estimation Challenges: Estimating $D_{\mathcal{H}}(\mu_{H_t|s_1} \| \mu_{H_t|s_2})$ is challenging due to the high dimensionality (d_{model}) of $\mathcal{H} \cong \mathbb{R}^{d_{\text{model}}}$ and the potentially complex geometry of the measures $\mu_{H_t|s}$. Standard techniques include:

- *Non-parametric methods*: Estimators based on k-nearest neighbor distances can approximate certain divergences like D_{KL} [21] or Rényi divergences. However, their statistical efficiency degrades rapidly with increasing dimension (curse of dimensionality), requiring a large number of samples h_t drawn from each conditional distribution.
- *Variational methods*: Techniques utilizing neural networks to estimate density ratios or variational bounds offer a potential alternative that may scale better with dimension. Examples include MINE for KL divergence [3] and methods for general f-divergences [13]. These rely on optimizing a neural network critic function $T : \mathcal{H} \to \mathbb{R}$ to approximate the divergence, e.g., via the Donsker-Varadhan representation for D_{KL}: $D_{\mathrm{KL}}(\mu\|\nu) = \sup_T(\mathbb{E}_\mu[T] - \log \mathbb{E}_\nu[e^T])$. These methods introduce optimization complexity and potential biases from the limited capacity of the critic network.

Estimation requires sampling contexts $\mathbf{w}_{<t}^{(i)} \sim P_{\mathrm{ctx}}(\cdot|s_1)$ and $\mathbf{w}_{<t}^{(j)} \sim P_{\mathrm{ctx}}(\cdot|s_2)$, computing the corresponding hidden states $h_t^{(i)} = (f_{\mathrm{bb}} \circ f_{\mathrm{emb}})(\mathbf{w}_{<t}^{(i)})$ and $h_t^{(j)} = (f_{\mathrm{bb}} \circ f_{\mathrm{emb}})(\mathbf{w}_{<t}^{(j)})$, and feeding these samples $\{h_t^{(i)}\}$ and $\{h_t^{(j)}\}$ into the chosen divergence estimator.

3.2 Metric 2: Categorical Mutual Information (Statistical Dependencies)

Goal: To measure the strength of statistical dependence between key random variables involved in the generation step, using the intrinsic definition of mutual information within the Markov Category framework.

Setup and Metrics: We use the categorical mutual information I_D (Equation (8)), which measures the D-divergence between a joint state and the product of its marginals.

1. **State-Prediction Dependence ($I_D(H_t; W_t)$)**: We aim to quantify the statistical dependence between the final hidden state H_t and the predicted next token W_t. This requires defining the joint state $p_{H_t,W_t} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H})) \otimes (\mathcal{V}, \mathcal{P}(\mathcal{V}))$, representing the joint distribution of (H_t, W_t) induced by the process $p_{W_{<t}} \to p_{H_t} \to p_{W_t}$. This state is obtained categorically by taking the state p_{H_t}, copying the $\mathcal{H}$ component using $\Delta_{\mathcal{H}} : \mathcal{H} \to \mathcal{H} \otimes \mathcal{H}$, and then applying the LM head kernel k_{head} only to the second component using the tensored identity $\mathrm{id}_{\mathcal{H}} \otimes k_{\mathrm{head}}$:

$$p_{H_t,W_t} \coloneqq (\mathrm{id}_{\mathcal{H}} \otimes k_{\mathrm{head}}) \circ \Delta_{\mathcal{H}} \circ p_{H_t}. \tag{14}$$

   The marginal states p_{H_t} and p_{W_t} can be recovered from p_{H_t,W_t} by discarding the other component using the unique maps $!_{\mathcal{V}} : \mathcal{V} \to I$ and $!_{\mathcal{H}} : \mathcal{H} \to I$: $p_{H_t} = (\mathrm{id}_{\mathcal{H}} \otimes !_{\mathcal{V}}) \circ p_{H_t,W_t}$ and $p_{W_t} = (!_{\mathcal{H}} \otimes \mathrm{id}_{\mathcal{V}}) \circ p_{H_t,W_t}$. The categorical mutual information between H_t and W_t is then defined as:

$$I_D(H_t; W_t) \coloneqq I_D(p_{H_t,W_t}) \equiv D_{\mathcal{H}\otimes\mathcal{V}}(p_{H_t,W_t} \quad \| \quad p_{H_t} \otimes p_{W_t}). \tag{15}$$

   This measures the deviation of the joint distribution from independence, according to the geometry induced by D. If $D = D_{\mathrm{KL}}$, this recovers the standard Shannon mutual information $I(H_t; W_t)$.

2. **Temporal State Dependence ($I_D(H_t; H_{t+1})$)**: To analyze the coherence between consecutive hidden states, we need to model the transition from H_t to H_{t+1}. This involves generating $W_t \sim k_{\mathrm{head}}(H_t, \cdot)$, updating the context $\mathbf{w}_{\leq t} = \mathbf{w}_{<t}W_t$, and then computing $H_{t+1} = (f_{\mathrm{bb}} \circ f_{\mathrm{emb}})(\mathbf{w}_{\leq t})$. This defines a complex transition kernel $k_{\mathrm{step}} : \mathcal{H} \to \mathcal{H}$ which implicitly

depends on the full history through H_t and the generated W_t. Assuming we average over the generation of W_t and the distribution p_{H_t}, we can define an effective transition kernel $\bar{k}_{\text{step}} : \mathcal{H} \to \mathcal{H}$. The joint state $p_{H_t,H_{t+1}} : I \to \mathcal{H} \otimes \mathcal{H}$ is constructed similarly to Equation (14):

$$p_{H_t,H_{t+1}} \coloneqq (\mathrm{id}_{\mathcal{H}} \otimes \bar{k}_{\text{step}}) \circ \Delta_{\mathcal{H}} \circ p_{H_t}. \tag{16}$$

The temporal statistical dependence is measured by:

$$I_D(H_t; H_{t+1}) \coloneqq I_D(p_{H_t,H_{t+1}}) \equiv D_{\mathcal{H}\otimes\mathcal{H}}(p_{H_t,H_{t+1}} \quad \| \quad p_{H_t} \otimes p_{H_{t+1}}), \tag{17}$$

where $p_{H_{t+1}} = (!_{\mathcal{H}} \otimes \mathrm{id}_{\mathcal{H}}) \circ p_{H_t,H_{t+1}} = \bar{k}_{\text{step}} \circ p_{H_t}$ is the marginal state at time $t+1$.

Interpretation: $I_D(H_t; W_t)$ quantifies the average amount of information (relative to divergence D) that the hidden state H_t provides about the immediately following token W_t. A high value suggests H_t strongly constrains the distribution over W_t, indicating high predictive relevance. $I_D(H_t; H_{t+1})$ measures the average statistical dependency between consecutive hidden states. A high value implies that the state H_{t+1} is highly predictable from H_t, suggesting the model maintains and evolves contextual information coherently over time. Low values might indicate information loss or abrupt changes in representation between time steps.

Estimation Challenges: Estimating I_D involving the high-dimensional continuous variable H_t (and potentially H_{t+1}) is difficult.

- For $I_D(H_t; W_t)$: W_t is discrete (finite vocabulary $\mathcal{V}$), while H_t is high-dimensional continuous. Mutual information estimation in such mixed settings is non-trivial. One could adapt general high-dimensional MI estimators like kNN-based methods (e.g., Kraskov-Stögbauer-Grassberger [11]) or variational methods (e.g., MINE [3]) by treating W_t appropriately (e.g., conditioning or embedding).
- For $I_D(H_t; H_{t+1})$: Both variables are high-dimensional and continuous. This poses the most significant estimation challenge, requiring robust high-dimensional MI estimators (kNN or variational) and potentially large sample sizes.

Estimation requires generating trajectories: sample $\mathbf{w}_{<t} \sim P_{\text{ctx}}$, compute $h_t = (f_{\text{bb}} \circ f_{\text{emb}})(\mathbf{w}_{<t})$, sample $w_t \sim k_{\text{head}}(h_t, \cdot)$, form $\mathbf{w}_{\leq t} = \mathbf{w}_{<t} w_t$, compute $h_{t+1} = (f_{\text{bb}} \circ f_{\text{emb}})(\mathbf{w}_{\leq t})$, and collect pairs (h_t, w_t) and (h_t, h_{t+1}) for input into the chosen MI estimator.

3.3 Metric 3: LM Head Categorical Entropy (Prediction Stochasticity)

Goal: To quantify the intrinsic stochasticity or uncertainty associated with the final prediction step, embodied by the LM head kernel $k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$.

Setup: We focus on the properties of the kernel k_{head} itself. The definition of categorical entropy (Equation (7)) involves comparing two processes that generate pairs of outputs in $Y \otimes Y$ from inputs in X, where $k : X \to Y$.

Metric: The Categorical Entropy of k_{head} is defined using Equation (7) with $X = \mathcal{H}$, $Y = \mathcal{V}$, and $k = k_{\text{head}}$:

$$\mathcal{H}_D(k_{\text{head}}) \coloneqq D_{\mathcal{V}\otimes\mathcal{V}}\left(\Delta_{\mathcal{V}} \circ k_{\text{head}} \quad \| \quad (k_{\text{head}} \otimes k_{\text{head}}) \circ \Delta_{\mathcal{H}}\right). \tag{18}$$

Let's analyze the two morphisms inside the divergence $D_{\mathcal{V}\otimes\mathcal{V}}(\cdot\|\cdot)$, which map $\mathcal{H} \to \mathcal{V} \otimes \mathcal{V}$:

- $k_1 = \Delta_{\mathcal{V}} \circ k_{\text{head}}$: For an input $h \in \mathcal{H}$, this first applies k_{head} to get a distribution $p_h(\cdot) = k_{\text{head}}(h, \cdot)$ on $\mathcal{V}$. Then, it applies the deterministic copy map $\Delta_{\mathcal{V}}(w, \cdot) = \delta_{(w,w)}$. The resulting kernel $k_1(h, \cdot)$ corresponds to sampling $w \sim p_h(\cdot)$ and then outputting the pair (w, w). The measure is $\sum_{w \in \mathcal{V}} p_h(w)\delta_{(w,w)}$ on $\mathcal{V} \otimes \mathcal{V}$.
- $k_2 = (k_{\text{head}} \otimes k_{\text{head}}) \circ \Delta_{\mathcal{H}}$: For an input $h \in \mathcal{H}$, this first applies the deterministic copy map $\Delta_{\mathcal{H}}(h, \cdot) = \delta_{(h,h)}$, producing the pair (h, h). Then, it applies k_{head} independently to each component via the tensor product kernel $(k_{\text{head}} \otimes k_{\text{head}})((h, h), \cdot) = k_{\text{head}}(h, \cdot) \otimes k_{\text{head}}(h, \cdot)$. The resulting kernel $k_2(h, \cdot)$ corresponds to sampling $w_1 \sim p_h(\cdot)$ and $w_2 \sim p_h(\cdot)$ independently and outputting the pair (w_1, w_2). The measure is $p_h \otimes p_h$ (the product measure) on $\mathcal{V} \otimes \mathcal{V}$.

The categorical entropy $\mathcal{H}_D(k_{\text{head}})$ thus measures the divergence between generating (W, W) where $W \sim p_h(\cdot)$ and generating (W_1, W_2) where $W_1, W_2 \overset{\text{i.i.d.}}{\sim} p_h(\cdot)$. This divergence

$D_{\mathcal{V}\otimes\mathcal{V}}(\sum_w p_h(w)\delta_{(w,w)} \| p_h \otimes p_h)$ quantifies how far p_h is from a point mass (Dirac measure), averaged appropriately over the input space $\mathcal{H}$. Note that $\mathcal{H}_D(k)$ as defined in [15] can be interpreted as a state on X (specifically $X = \mathcal{H}$ here) whose value at h relates to this divergence. A common practical approach is to consider the average divergence with respect to the input distribution p_{H_t}:

$$\bar{\mathcal{H}}_D(k_{\text{head}}; p_{H_t}) \coloneqq \mathbb{E}_{h \sim p_{H_t}} \left[D_{\mathcal{V}\otimes\mathcal{V}} \left(\sum_{w \in \mathcal{V}} k_{\text{head}}(h, \{w\})\delta_{(w,w)} \quad \| \quad k_{\text{head}}(h, \cdot) \otimes k_{\text{head}}(h, \cdot) \right) \right]. \tag{19}$$

Interpretation: This metric measures the intrinsic conditional stochasticity of the LM head mapping. If k_{head} were deterministic (i.e., for each h, it mapped to a single specific w_h, so $p_h = \delta_{w_h}$), then both measures inside the divergence would be $\delta_{(w_h,w_h)}$, and the entropy would be $D(\delta_{(w_h,w_h)} \| \delta_{(w_h,w_h)}) = 0$. A higher value of $\mathcal{H}_D(k_{\text{head}})$ indicates greater average uncertainty or "spread" in the output distribution $p_h = k_{\text{head}}(h, \cdot)$, meaning the kernel is inherently more stochastic. It quantifies how far the prediction process is from a deterministic assignment, measured in the geometry of $\mathcal{V} \otimes \mathcal{V}$ induced by D.

For the specific case $D = D_{\text{KL}}$, the inner divergence relates closely to the Shannon entropy of p_h. The average categorical entropy $\bar{\mathcal{H}}_{D_{\text{KL}}}(k_{\text{head}}; p_{H_t})$ provides a measure akin to the average conditional Shannon entropy:

$$\mathbb{E}_{h \sim p_{H_t}}[H(k_{\text{head}}(h, \cdot))] = \mathbb{E}_{h \sim p_{H_t}} \left[-\sum_{w \in \mathcal{V}} k_{\text{head}}(h, \{w\}) \log k_{\text{head}}(h, \{w\}) \right]. \tag{20}$$

While not identical, both measures capture the average uncertainty in the next-token prediction given the hidden state.

Estimation: Estimating the average categorical entropy (Equation (19)) involves averaging over samples $h_t \sim p_{H_t}$. For each sampled h_t: 1. Compute the output probability vector $p_{h_t} = [k_{\text{head}}(h_t, \{w\})]_{w \in \mathcal{V}}$. 2. Construct the two required probability measures on the finite space $\mathcal{V} \times \mathcal{V}$: $\mu_1 = \sum_w p_{h_t}(w)\delta_{(w,w)}$ and $\mu_2 = p_{h_t} \otimes p_{h_t}$. 3. Compute the divergence $D_{\mathcal{V}\otimes\mathcal{V}}(\mu_1 \| \mu_2)$. Since the space is finite, this calculation is often straightforward (e.g., for KL divergence:

$\sum_{(w_1,w_2)} \mu_1(w_1, w_2) \log(\mu_1(w_1, w_2)/\mu_2(w_1, w_2))$). 4. Average these divergence values over many samples of h_t obtained by sampling contexts $\mathbf{w}_{<t} \sim P_{\text{ctx}}$ and applying $k_{\text{emb}}, k_{\text{bb}}$. Estimating the

average Shannon entropy (Equation (20)) follows a similar Monte Carlo approach, calculating $H(p_{h_t})$ in step 3.

3.4 Metric 4: Information Flow Bounds via Data Processing Inequality

Goal: To leverage the fundamental Data Processing Inequality (DPI), inherent in the Markov category **Stoch** and satisfied by I_D, to establish bounds on how much information about a context property S can propagate through the processing chain to the final output token W_t.

Setup: Let S be a property of the context $\mathbf{w}_{<t}$ as in Metric 1. The sequence of transformations $S \to \mathbf{w}_{<t} \to E_{<t} \to H_t \to W_t$ forms a Markov chain, provided we consider the joint distribution $P(s, \mathbf{w}_{<t}, e_{<t}, h_t, w_t)$ induced by the process. Since k_{emb} and k_{bb} are deterministic functions of $\mathbf{w}_{<t}$, H_t is a function of $\mathbf{w}_{<t}$. Furthermore, k_{head} generates W_t based only on H_t. Thus, we have the Markov chain structure $S \to H_t \to W_t$. This means the conditional distribution of W_t given H_t and S depends only on H_t: $P(W_t|H_t, S) = P(W_t|H_t)$.

Consider the joint distribution of (S, H_t), represented by the state $p_{S,H_t} : I \to S \otimes (\mathcal{H}, \mathscr{B}(\mathcal{H}))$ (assuming S takes values in a measurable space, also denoted S). Similarly, let $p_{S,W_t} : I \to S \otimes (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ be the joint state of (S, W_t). Crucially, the state p_{S,W_t} can be obtained from p_{S,H_t} by applying the kernel $\mathrm{id}_S \otimes k_{\text{head}} : S \otimes (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to S \otimes (\mathcal{V}, \mathcal{P}(\mathcal{V}))$, which acts as k_{head} on the $\mathcal{H}$ component while leaving the S component unchanged:

$$p_{S,W_t} = (\mathrm{id}_S \otimes k_{\text{head}}) \circ p_{S,H_t}. \tag{21}$$

Theorem 3.1 (Categorical Information Flow Bound)**.** *: Let $I_D(S; X) := D_{S\otimes X}(p_{S,X} \| p_S \otimes p_X)$ denote the categorical mutual information between S and $X \in \{H_t, W_t\}$, where $p_{S,X}$ is the joint state and p_S, p_X are the corresponding marginal states. The Data Processing Inequality for the divergence D, when applied to the definition of I_D and the Markov kernel $\mathrm{id}_S \otimes k_{\text{head}}$ processing p_{S,H_t} to p_{S,W_t}, implies:*

$$I_D(S; H_t) \geq I_D(S; W_t). \tag{22}$$

Proof sketch: The DPI states that for any kernel $k : A \to B$ and states $p, q : I \to A$, we have $D_B(k \circ p \| k \circ q) \leq D_A(p\|q)$. Apply this with $A = S \otimes \mathcal{H}$, $B = S \otimes \mathcal{V}$, $k = \mathrm{id}_S \otimes k_{\text{head}}$, $p = p_{S,H_t}$, and $q = p_S \otimes p_{H_t}$. Then $k \circ p = p_{S,W_t}$ and $k \circ q = (\mathrm{id}_S \otimes k_{\text{head}}) \circ (p_S \otimes p_{H_t}) = p_S \otimes (k_{\text{head}} \circ p_{H_t}) = p_S \otimes p_{W_t}$. The inequality $D_{S\otimes\mathcal{V}}(p_{S,W_t} \| p_S \otimes p_{W_t}) \leq D_{S\otimes\mathcal{H}}(p_{S,H_t} \| p_S \otimes p_{H_t})$ directly yields $I_D(S; W_t) \leq I_D(S; H_t)$.

Interpretation: This fundamental inequality asserts that the amount of statistical information (measured by I_D) that the next token W_t carries about the context property S cannot exceed the amount of information about S that is already encoded in the intermediate hidden representation H_t. The final stochastic step $k_{\text{head}} : H_t \to W_t$ can only preserve or lose information about S; it cannot create it. The difference $I_D(S; H_t) - I_D(S; W_t) \geq 0$ quantifies the information about S that is present in the representation H_t but is "lost" or not utilized in the immediate prediction of W_t. This loss could be due to the inherent stochasticity of k_{head} (as measured by $\mathcal{H}_D(k_{\text{head}})$) or because the mapping discards aspects of H_t relevant to S but not relevant for predicting W_t. This unused information might still be crucial for predicting subsequent tokens ($W_{t+1}, \ldots$).

Estimation Challenges: Requires estimating two mutual information quantities:

- $I_D(S; W_t)$: If S is discrete or low-dimensional, this involves MI between S and the discrete variable W_t. This is often the more tractable quantity to estimate, potentially using direct frequency counts or simple estimators if S has few categories.
- $I_D(S; H_t)$: This involves MI between the context property S and the high-dimensional continuous hidden state H_t. This estimation faces the same challenges as $I_D(H_t; W_t)$ and $I_D(H_t; H_{t+1})$, requiring robust high-dimensional MI estimators (kNN or variational) sensitive to the specific structure of S (discrete vs. continuous).

Estimation relies on obtaining samples of (s, h_t) and (s, w_t) pairs. This is done by sampling contexts $\mathbf{w}_{<t}$ associated with property value s (i.e., from $P_{\text{ctx}}(\cdot|s)$), running the AR generation step ($k_{\text{emb}}, k_{\text{bb}}, k_{\text{head}}$) to get h_t and w_t, and collecting these samples for input into the chosen MI estimators. Comparing the estimated values provides an empirical check on the information flow bottleneck at the LM head stage.

4 Pretraining Objective, Compression, and Categorical Entropy

A central question surrounding large language models is why the seemingly simple auto-regressive objective of next-token prediction, trained via minimizing cross-entropy loss, yields such powerful and versatile capabilities, often exhibiting behaviors associated with understanding and reasoning. The framework of Markov Categories and categorical entropy provides a lens through which to interpret this phenomenon, connecting it to fundamental ideas about compression and information.

The standard pretraining objective for an AR LM is to minimize the negative log-likelihood (NLL) of the next token w_t given the preceding context $\mathbf{w}_{<t}$, averaged over a large text corpus. Let $P_{\text{data}}(\mathbf{w}_{<t}, w_t)$ be the empirical joint distribution observed in the training data, inducing the conditional distribution $P_{\text{data}}(w_t|\mathbf{w}_{<t})$. The model parameterizes a conditional distribution $P_\theta(w_t|\mathbf{w}_{<t})$. The loss function is:

$$L_{\text{CE}}(\theta) = -\mathbb{E}_{(\mathbf{w}_{<t}, w_t)\sim P_{\text{data}}}[\log P_\theta(w_t|\mathbf{w}_{<t})] \tag{23}$$

Minimizing this cross-entropy is equivalent to minimizing the average Kullback-Leibler (KL) divergence between the empirical conditional distribution and the model's conditional distribution:

$$\arg\min_\theta L_{\text{CE}}(\theta) = \arg\min_\theta \mathbb{E}_{\mathbf{w}_{<t}\sim P_{\text{data}}}[D_{\text{KL}}(P_{\text{data}}(\cdot|\mathbf{w}_{<t})\|P_\theta(\cdot|\mathbf{w}_{<t}))] + H(W_t|W_{<t})_{\text{data}} \tag{24}$$

where $H(W_t|W_{<t})_{\text{data}}$ is the conditional Shannon entropy of the data-generating process, which is independent of θ.

Let us represent the true (but unknown) data-generating process as a Markov kernel $k_{\text{data}} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$, such that $k_{\text{data}}(\mathbf{w}_{<t}, \cdot)$ corresponds to the measure $P_{\text{data}}(\cdot|\mathbf{w}_{<t})$. Similarly, the parameterized model corresponds to the kernel $k_{\text{gen},\theta} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$, defined in Equation (2). The training objective Equation (24) can be rewritten using these kernels:

$$\min_\theta \mathbb{E}_{\mathbf{w}_{<t}\sim p_{W_{<t}}}[D_{\text{KL}}(k_{\text{data}}(\mathbf{w}_{<t}, \cdot)\|k_{\text{gen},\theta}(\mathbf{w}_{<t}, \cdot))] \tag{25}$$

where $p_{W_{<t}}$ is the marginal distribution over contexts derived from P_{data}.

Assuming the model class is sufficiently expressive and the optimization is successful, the cross-entropy loss converges towards its global minimum. This minimum value is achieved when

$k_{\text{gen},\theta}(\mathbf{w}_{<t}, \cdot) = k_{\text{data}}(\mathbf{w}_{<t}, \cdot)$ for almost every context $\mathbf{w}_{<t}$ (with respect to $p_{W_{<t}}$). At this point, the model has effectively learned to replicate the true conditional probabilities of the language source, as observed in the data.

The connection to compression arises from Shannon's source coding theorem. The theoretical minimum average code length required to encode the next token w_t, given the context $\mathbf{w}_{<t}$, is precisely the conditional entropy $H(W_t|W_{<t})_{\text{data}}$. The cross-entropy loss $L_{CE}(\theta)$ achieved by the model represents the average code length obtained when using the model's distribution P_θ as the basis for an encoding scheme (e.g., arithmetic coding). Therefore, minimizing the cross-entropy loss is directly equivalent to finding a model P_θ that provides the most efficient compression of the training data, according to the principles of information theory. The minimal achievable loss (the data entropy) quantifies the inherent unpredictability or randomness in the language sequence, even given perfect knowledge of the preceding context.

The hypothesis that "compression is intelligence" posits that the ability to significantly compress data (i.e., find a compact description or generative model) requires uncovering and exploiting underlying structures, regularities, and causal relationships within the data. For natural language, these structures encompass grammar, semantics, factual knowledge, discourse coherence, and potentially even rudimentary forms of reasoning, all of which are necessary to accurately predict upcoming tokens. By optimizing for compression (i.e., minimizing cross-entropy), the AR LM is forced to internalize these complex structures within its parameters θ and internal computations represented by the kernels $k_{\text{emb}}, k_{\text{bb}}, k_{\text{head}}$. The emergent capabilities of LLMs can thus be viewed as a byproduct of achieving high compression rates on vast amounts of text data.

Within the Markov Category framework, the learned kernel $k_{\text{gen},\theta}$ itself represents the compressed knowledge extracted from the data. Its efficacy is measured by how closely it approximates k_{data} via the objective in Equation (25). If the global minimum is reached ($k_{\text{gen},\theta} \approx k_{\text{data}}$), we can consider the meaning of this compression through the lens of categorical entropy, $\mathcal{H}_D(k)$ (Equation (7)), where D is a chosen divergence.

$$\mathcal{H}_D(k) \coloneqq D_{Y \otimes Y} \left(\Delta_Y \circ k \quad \| \quad (k \otimes k) \circ \Delta_X \right) \tag{26}$$

This quantity measures the intrinsic stochasticity of the kernel k, comparing the deterministic copying of an output with the independent generation of two outputs from the same input. If the model converges such that $k_{\text{gen},\theta} \approx k_{\text{data}}$, then the categorical entropy of the learned kernel will approximate the categorical entropy of the true data-generating process:

$$\mathcal{H}_D(k_{\text{gen},\theta}) \approx \mathcal{H}_D(k_{\text{data}}) \tag{27}$$

The quantity $\mathcal{H}_D(k_{\text{data}})$ represents the fundamental, irreducible conditional stochasticity inherent in the language source itself, quantified relative to the divergence D. For instance, if $D = D_{\text{KL}}$, it relates to the average uncertainty (measured by D_{KL}) in predicting two subsequent tokens independently versus predicting one and copying it, given the same context.

Therefore, when the pretraining loss converges to its global minimum, the compression achieved means the model $k_{\text{gen},\theta}$ has not only matched the predictive accuracy of the true process k_{data} (minimizing KL divergence point-wise on average) but has also captured its intrinsic stochastic nature, as measured by the categorical entropy $\mathcal{H}_D$. The parameters θ and the corresponding compositional structure $k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ constitute the compressed representation of the predictive

rules of the language. The effectiveness of this compression, forced by the next-token prediction objective, necessitates the learning of linguistic and world knowledge, thereby leading to the powerful generalization capabilities observed in large language models.

5 Information Geometry of Representation and Prediction Spaces

The Markov Category framework, particularly $(\mathbf{Stoch}, D)$ enriched with a divergence like D_{KL}, provides a natural bridge to Information Geometry [1, 16]. This allows for a geometric analysis of the spaces involved in AR language modeling, particularly the representation space $\mathcal{H}$ and the space of next-token distributions $\mathcal{P}(\mathcal{V})$.

The space $\mathcal{P}(\mathcal{V})$ of probability distributions over the finite vocabulary $\mathcal{V}$ forms a $(|\mathcal{V}|-1)$-dimensional simplex $\Delta^{|\mathcal{V}|-1}$. This space possesses a well-defined Riemannian geometry induced by the Fisher-Rao information metric g^{FR}, whose components in a local coordinate system $\xi = (\xi_1, \ldots, \xi_{|\mathcal{V}|-1})$ for a distribution $p_\xi \in \mathcal{P}(\mathcal{V})$ are given by:

$$g^{\mathrm{FR}}_{ij}(\xi) = \sum_{w \in \mathcal{V}} p_\xi(w) \frac{\partial \log p_\xi(w)}{\partial \xi_i} \frac{\partial \log p_\xi(w)}{\partial \xi_j} = \mathbb{E}_{W \sim p_\xi}\left[\frac{\partial \log p_\xi(W)}{\partial \xi_i} \frac{\partial \log p_\xi(W)}{\partial \xi_j} \right]. \tag{28}$$

This metric quantifies the local distinguishability between nearby probability distributions, measuring the distance in terms of expected squared log-likelihood ratio gradients. The geometry of $\mathcal{P}(\mathcal{V})$ also includes dual affine connections ($\pm\alpha$-connections) related to the KL divergence, providing a richer dually flat structure [1].

The LM Head kernel $k_{\mathrm{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ defines a deterministic mapping from a hidden state $h \in \mathcal{H} \cong \mathbb{R}^{d_{\mathrm{model}}}$ to a probability distribution $p_h := k_{\mathrm{head}}(h, \cdot) \in \mathcal{P}(\mathcal{V})$. This mapping, let's call the underlying function $g_{\mathrm{head}} : \mathcal{H} \to \mathcal{P}(\mathcal{V})$, allows us to pull back the geometric structure from $\mathcal{P}(\mathcal{V})$ onto the representation space $\mathcal{H}$.

Specifically, the Fisher-Rao metric g^{FR} on $\mathcal{P}(\mathcal{V})$ induces a (generally degenerate) Riemannian metric tensor $g^* = g^*_{\mathrm{head}} g^{\mathrm{FR}}$ on $\mathcal{H}$. At a point $h \in \mathcal{H}$, the components of this pullback metric are given by:

$$g^*_{ab}(h) = \sum_{i,j} g^{\mathrm{FR}}_{ij}(g_{\mathrm{head}}(h)) \frac{\partial (g_{\mathrm{head}}(h))_i}{\partial h_a} \frac{\partial (g_{\mathrm{head}}(h))_j}{\partial h_b}, \quad a, b \in \{1, \ldots, d_{\mathrm{model}}\}, \tag{29}$$

where h_a, h_b are coordinates of $h \in \mathcal{H}$, and $(g_{\mathrm{head}}(h))_i, (g_{\mathrm{head}}(h))_j$ represent local coordinates of the output distribution $p_h \in \mathcal{P}(\mathcal{V})$ (e.g., probabilities of specific tokens, possibly excluding one due to the sum-to-one constraint). The term $\frac{\partial (g_{\mathrm{head}}(h))_i}{\partial h_a}$ is the Jacobian of the LM head map g_{head} evaluated at h.

Interpretation of the Pullback Metric g^*:

- **Sensitivity Analysis:** The quadratic form associated with $g^*(h)$, namely $g^*(h)(v, v)$ for a tangent vector $v \in T_h\mathcal{H} \cong \mathcal{H}$, measures the infinitesimal squared distance (according to g^{FR}) between the output distributions p_h and $p_{h+\epsilon v}$ for small ϵ. It quantifies the sensitivity of the model's prediction p_h to perturbations of the hidden state h in the direction v. Directions v with large $g^*(h)(v, v)$ correspond to changes in h that significantly alter the output distribution's geometry.

- **Degeneracy and Rank:** Since the dimension of $\mathcal{P}(\mathcal{V})$ is $|\mathcal{V}| - 1$, the rank of the pullback metric $g^*(h)$ is at most $|\mathcal{V}| - 1$. Given that typically $d_{\text{model}} \gg |\mathcal{V}|$, $g^*(h)$ is highly degenerate. Its null space, $\ker(g^*(h)) = \{v \in T_h\mathcal{H} \mid g^*(h)(v, w) = 0 \text{ for all } w \in T_h\mathcal{H}\}$, consists of directions v in $\mathcal{H}$ such that infinitesimal movements along v do not change the output distribution p_h at the first order, according to the Fisher-Rao metric. These directions represent representational changes irrelevant to the immediate next-token prediction.
- **Spectrum:** The eigenvalues and eigenvectors of $g^*(h)$ (restricted to its support, the orthogonal complement of the null space) reveal the principal directions of sensitivity in the representation space $\mathcal{H}$ with respect to the prediction task. Directions corresponding to large eigenvalues are those where small changes in h induce large changes (geometrically measured by g^{FR}) in the predicted distribution.

This geometric perspective provides a rigorous way to analyze the functional geometry of the representation space $\mathcal{H}$ as shaped by the downstream prediction task defined by k_{head}. Investigating how g^* varies across regions of $\mathcal{H}$ populated by the hidden state distribution p_{H_t} (Section 3) could reveal how the model allocates representational sensitivity. For example, we could compute the average metric $\bar{g}^* = \mathbb{E}_{h\sim p_{H_t}}[g^*(h)]$ or study the properties of the manifold $(\mathcal{H}, g^*)$ near typical representations h. Furthermore, concepts like the geometric volume element $\sqrt{\det g^*(h)}$ (on the support) or the Ricci curvature derived from g^* could offer novel insights into the learned representation manifold, potentially connecting to generalization properties or the complexity of the learned function, perhaps drawing parallels with singular learning theory [23] which studies the geometry of parameter spaces near singularities. The dual connections on $\mathcal{P}(\mathcal{V})$ can also be pulled back via g_{head}, providing further geometric tools (dual metrics, curvatures) for analyzing $\mathcal{H}$.

6 Next-Token Prediction as Implicit Structure Learning in $\mathcal{H}$

A central question in the study of large language models is why the objective of minimizing the negative log-likelihood (NLL) of the next token (Equations (24) and (25)) induces internal representations $h_t \in \mathcal{H}$ that capture rich semantic, syntactic, and contextual information. While AR models lack the explicit positive/negative pairing structure of typical contrastive learning setups, we posit that the NLL objective itself implicitly imposes a form of structural constraint on the learned representations, akin to spectral methods explored in self-supervised learning [24, 25].

Let $f_{\text{enc}} : \mathcal{V}^* \to \mathcal{H}$ denote the deterministic encoder mapping a context sequence $x = \mathbf{w}_{<t}$ to its hidden representation $h_x = f_{\text{enc}}(x)$, implemented by $k_{\text{bb}} \circ k_{\text{emb}}$. Let $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathcal{V})$ be the deterministic mapping from the hidden state to the next-token distribution, corresponding to the LM head kernel k_{head}, i.e., $p_\theta(\cdot|x) = g_{\text{head}}(h_x)$. The training objective is to minimize the expected KL divergence:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim p_{W_{<t}}}[D_{\text{KL}}(P_{\text{data}}(\cdot|x) \,\|\, g_{\text{head}}(f_{\text{enc}}(x)))] \tag{30}$$

where $p_{W_{<t}}$ is the distribution over contexts and $P_{\text{data}}(\cdot|x)$ is the true conditional distribution of the next token given context x.

Successful optimization ensures that $g_{\text{head}}(h_x)$ closely approximates $P_{\text{data}}(\cdot|x)$ for most contexts x. This implies an implicit contrastive pressure: if two contexts x and x' yield significantly different true next-token distributions $P_{\text{data}}(\cdot|x)$ and $P_{\text{data}}(\cdot|x')$, the model must map them to representations

h_x and $h_{x'}$ such that $g_{\text{head}}(h_x)$ and $g_{\text{head}}(h_{x'})$ are correspondingly distinct within the probability simplex $\mathcal{P}(\mathcal{V})$.

We can quantify the dissimilarity between target distributions using metrics like the Hellinger distance d_H or the symmetric KL divergence $d_{SKL}(p\|q) = \frac{1}{2}(D_{\text{KL}}(p\|q) + D_{\text{KL}}(q\|p))$. Since $d_H^2(p,q) \leq \frac{1}{2}D_{\text{KL}}(p\|q)$ (Pinsker's inequality, adjusted for base e) and the triangle inequality holds for d_H, achieving low average KL loss $\mathcal{L}(\theta)$ implies that, on average:

$$d_H(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'})) \approx d_H(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x')). \tag{31}$$

Therefore, contexts x, x' that are "predictively dissimilar" under P_{data} (large $d_H(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$) must be mapped to representations $h_x, h_{x'}$ whose images under g_{head} are far apart in $\mathcal{P}(\mathcal{V})$. If g_{head} exhibits sufficient local sensitivity (i.e., the pullback metric g^* discussed in Section 5 is non-degenerate in relevant directions), this necessitates a separation between h_x and $h_{x'}$ in the representation space $\mathcal{H}$. The NLL objective implicitly pushes representations apart based on the predictive dissimilarity of their corresponding contexts.

This perspective invites analogies with spectral methods in representation learning. Let us define a measure of *predictive similarity* between contexts x and x'.

Definition 6.1 (Predictive Similarity Kernel)**.** *Let $p_x := P_{data}(\cdot|x)$. Define a similarity function $K : \mathcal{V}^* \times \mathcal{V}^* \to \mathbb{R}_{\geq 0}$ based on the overlap or proximity of the true next-token distributions. Examples include:*

- ***Bhattacharyya Coefficient Kernel**: $K_{BC}(x, x') := BC(p_x, p_{x'}) = \sum_{w \in \mathcal{V}} \sqrt{p_x(w)p_{x'}(w)}$.*
- ***Hellinger-based Kernel**: $K_H(x, x') := \exp(-\beta d_H^2(p_x, p_{x'}))$ for some scale $\beta > 0$.*
- ***Expected Likelihood Kernel (ELK)**: $K_{ELK}(x, x') := \frac{1}{2}(\mathbb{E}_{W \sim p_x}[p_{x'}(W)] + \mathbb{E}_{W \sim p_{x'}}[p_x(W)])$.*

High values of $K(x, x')$ indicate that contexts x and x' lead to similar predictions under the true data distribution.

6.1 Analogy to Spectral Clustering

Following the perspective of [25], we can conceptualize an infinite graph where nodes are contexts $x \in \mathcal{V}^*$ and edge weights are given by the predictive similarity kernel $K(x, x')$. Spectral clustering on such a graph aims to find a low-dimensional embedding (representation) $\psi : \mathcal{V}^* \to \mathbb{R}^d$ such that similar contexts (high $K(x, x')$) are mapped close together. This is often achieved by finding eigenfunctions of the graph Laplacian associated with K.

Let μ_{ctx} be the measure on $\mathcal{V}^*$ corresponding to $p_{W_{<t}}$. Define the degree function $d(x) = \int_{\mathcal{V}^*} K(x, x')\, \mu_{ctx}(\mathrm{d}x')$. The normalized graph Laplacian $\mathcal{L}$ acts on functions $\phi : \mathcal{V}^* \to \mathbb{R}$ as:

$$(\mathcal{L}\phi)(x) = \phi(x) - \int_{\mathcal{V}^*} \frac{K(x, x')}{\sqrt{d(x)d(x')}} \phi(x')\, \mu_{ctx}(\mathrm{d}x'). \tag{32}$$

Minimizing the NLL objective $\mathcal{L}(\theta)$ (Equation (30)) implicitly encourages the learned encoder f_{enc} to respect this similarity structure. If $K(x, x')$ is high, then $p_x \approx p_{x'}$. Low NLL requires $g_{\text{head}}(h_x) \approx p_x$ and $g_{\text{head}}(h_{x'}) \approx p_{x'}$, implying $g_{\text{head}}(h_x) \approx g_{\text{head}}(h_{x'})$. As argued via Equation (31)

and the sensitivity of g_{head}, this pushes h_x and $h_{x'}$ closer in $\mathcal{H}$. Conversely, if $K(x, x')$ is low, the objective forces $g_{\text{head}}(h_x)$ and $g_{\text{head}}(h_{x'})$ apart, thus separating h_x and $h_{x'}$.

This behavior aligns qualitatively with minimizing the Dirichlet energy associated with the Laplacian:

$$\mathcal{E}(\phi) = \frac{1}{2} \iint_{\mathcal{V}^* \times \mathcal{V}^*} K(x, x')(\phi(x) - \phi(x'))^2 \, \mu_{ctx}(\mathrm{d}x)\mu_{ctx}(\mathrm{d}x'), \tag{33}$$

whose minimizers (under constraints) are the eigenfunctions of $\mathcal{L}$. While NLL minimization does not explicitly optimize Equation (33) for the components of f_{enc}, the pressures it exerts on the representations $h_x = f_{\text{enc}}(x)$ push them towards configurations favored by spectral clustering on the predictive similarity graph. The structure learned in $\mathcal{H}$ reflects clusters of contexts that are predictively similar according to P_{data}.

6.2 Analogy to Spectral Contrastive Representation Learning

Alternatively, inspired by [24], we can consider an integral operator defined using the predictive similarity kernel. Let $\mu = p_{H_t}$ be the distribution of hidden states h_x induced by f_{enc} and $p_{W_{<t}}$. Define the operator M acting on functions $\phi \in L^2(\mathcal{H}, \mu)$:

$$(M\phi)(h_x) = \mathbb{E}_{x' \sim p_{W_{<t}}}[K(x, x')\phi(h_{x'})] = \int_{\mathcal{V}^*} K(x, x')\phi(f_{\text{enc}}(x')) \, p_{W_{<t}}(\mathrm{d}x'). \tag{34}$$

This operator averages the function ϕ over representations $h_{x'}$ weighted by their predictive similarity $K(x, x')$ to the reference context x.

In [24], it was shown that spectral contrastive loss encourages representations to align with the top eigenspace of a population augmentation operator P_{aa}, which averages representations over augmentations known to preserve class labels. In the AR setting, there are no explicit augmentations. However, we can view contexts x' with high predictive similarity $K(x, x')$ to x as being "semantically related" from the perspective of the next-token prediction task.

Minimizing the NLL objective forces h_x and $h_{x'}$ to be "compatible" under g_{head} when $K(x, x')$ is large. If we assume the learned representations h_x capture the underlying structure smoothly, optimizing NLL might implicitly favor representations that are eigenvectors of M with large eigenvalues. Such eigenvectors represent functions ϕ on $\mathcal{H}$ that co-vary positively with the predictive similarity structure; $\phi(h_x)$ tends to be large when h_x corresponds to contexts x that are predictively similar to many other contexts x'. While M differs from the augmentation operator P_{aa} in [24], it similarly captures a notion of task-relevant similarity. The NLL objective, by forcing predictions to match P_{data}, implicitly encourages alignment with the structure revealed by the eigenspectrum of the predictive similarity operator M.

6.3 Discussion

The connections drawn here are analogies, highlighting how the pressures of the NLL objective qualitatively resemble those of spectral methods operating on a graph defined by predictive similarity. Rigorously proving that f_{enc} converges to eigenfunctions of $\mathcal{L}$ or M directly from minimizing $\mathcal{L}(\theta)$ (Equation (30)) remains an open challenge. Key difficulties include:

1. The NLL objective is a point-wise loss (minimizing KL for each x), not directly a pairwise or spectral objective like Equation (33) or those in [24, 25].
2. The relationship between representation distance

   $\|h_x - h_{x'}\|$ and prediction distance $d(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ is mediated by the potentially complex, non-linear geometry induced by the LM head g_{head} (characterized by g^* in Section 5).
3. The structure is learned relative to the *true* data distribution $P_{\text{data}}(\cdot|x)$, which is only accessed through finite samples.

Despite these challenges, this perspective suggests that the success of next-token prediction in inducing structured representations is not entirely mysterious. By forcing the model to internalize the complex conditional dependencies $P_{\text{data}}(\cdot|x)$ of language, the NLL objective implicitly sculpts the representation space $\mathcal{H}$ to reflect the underlying predictive similarities and dissimilarities between contexts, mirroring principles from spectral graph theory and representation learning. The learned geometry of $\mathcal{H}$ is thus deeply tied to the predictive structure of the language itself.

7 Limitations and Future Directions

Limitations:

- **Static vs. Dynamic Analysis**: Our primary focus has been on the single-step generation kernel $k_{\text{gen},\theta} : \mathcal{V}^* \to \mathcal{V}$. A full analysis of sequence generation requires understanding the dynamics over multiple steps, involving the iterated composition of kernels and the feedback loop where the generated token modifies the subsequent context. Modeling this full dynamic process within the MC framework, possibly using techniques for iterated systems or graphical models, presents further theoretical and computational challenges.
- **Sufficiency of H_t**: The analysis often implicitly assumes H_t (typically the final layer's output at the last position) captures all relevant context for the next token prediction. While Transformers have access to all past token representations via attention, the final h_t acts as an information bottleneck before the LM head. The extent to which h_t is a sufficient statistic for the future W_t given the past $\mathbf{w}_{<t}$ is crucial and may not hold perfectly. Analyzing information flow directly from the input sequence $\mathbf{w}_{<t}$ to W_t versus via H_t could quantify this bottleneck.
- **Theoretical Gaps in Representation Learning Analogies**: The connections drawn to spectral clustering and spectral contrastive learning in Section 6 are currently qualitative analogies. Rigorously proving convergence of representations learned via NLL minimization to specific structures (e.g., eigenspaces of predictive similarity operators) requires bridging the gap between point-wise loss minimization and pairwise/spectral objectives, potentially involving assumptions about the data distribution and model architecture.
- **Choice of Divergence D**: The framework relies on choosing a specific divergence D. Different divergences capture different aspects of distributional dissimilarity (e.g., KL vs. TV vs. Rényi). The choice impacts the interpretation and values of the categorical entropy and MI metrics. Understanding the implications of different divergence choices is necessary.

Future Directions:

1. **Formalizing Spectral Connections**: Further investigate the theoretical links between the NLL objective and spectral properties of operators defined on the context or representation space (Section 6). Can conditions be identified under which NLL minimization provably aligns representations with specific eigenfunctions relevant to prediction? This might involve analyzing the loss landscape geometry or using tools from optimal transport or functional analysis.
2. **MC-Inspired Model Design**: Can the insights gained from this categorical and information-geometric analysis inform the design of new LM architectures or training objectives? For example, could explicit regularization based on controlling $I_D(H_t; W_t)$, $\mathcal{H}_D(k_{\text{head}})$, the rank of the pullback metric g^* (Equation (29)), or the information flow gap $I_D(S; H_t) - I_D(S; W_t)$ lead to models with improved interpretability, sample efficiency, compositional generalization, or robustness? Could MC principles like explicit compositionality guide the design of more modular or verifiable LM architectures?
3. **Multi-Step Dynamics and Control**: Extend the framework to analyze multi-step generation dynamics. This could involve studying the properties of composed kernels $k^{(n)}_{\text{gen},\theta}$ or using tools from dynamical systems theory adapted to the MC setting. Furthermore, exploring connections to optimal control theory within MCs might offer perspectives on guided text generation or steering model behavior.

8 Conclusion

In this work, we introduced a rigorous mathematical framework for analyzing the core autoregressive generation step in language models, leveraging the expressive power of Markov Categories (MCs). By explicitly modeling the context-to-prediction mapping as a composition of Markov kernels $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ within the canonical category **Stoch**, we provided a foundation for compositional analysis that respects the probabilistic nature of the process. The enrichment of **Stoch** with a statistical divergence D, and the subsequent use of categorical entropy $\mathcal{H}_D$ and mutual information I_D as defined by Perrone [15], enabled the development of novel metrics. These metrics offer principled ways to quantify: the ability of hidden states H_t to encode contextual information (RepDiv$_D$ via $D(p_{H_t|s_1} \| p_{H_t|s_2})$); the statistical coupling between representations and predictions ($I_D(H_t; W_t)$) and across time steps ($I_D(H_t; H_{t+1})$); the inherent stochasticity of the prediction head ($\mathcal{H}_D(k_{\text{head}})$); and fundamental information flow constraints derived from the Data Processing Inequality ($I_D(S; H_t) \geq I_D(S; W_t)$).

Furthermore, we explored the connections between this framework and related theoretical areas. The information geometry perspective (Section 5) reveals how the prediction task induces a specific geometry (via the pullback Fisher-Rao metric g^*) on the representation space $\mathcal{H}$, quantifying the sensitivity of predictions to representational changes. Our analysis in Sections 4 and 6 suggests that the standard cross-entropy training objective, interpreted as minimizing the KL divergence between the model kernel $k_{\text{gen},\theta}$ and the data kernel k_{data}, implicitly forces the model to perform a type of compression that captures the intrinsic stochasticity of language (related to $\mathcal{H}_D(k_{\text{data}})$). Moreover, we proposed that this objective implicitly sculpts the representation space $\mathcal{H}$ by separating representations based on the predictive dissimilarity of their corresponding contexts, drawing analogies to spectral methods operating on graphs defined by predictive similarity kernels.

This Markov Categorical approach provides a unified lens, integrating concepts from category theory, probability theory, information theory, and information geometry, to move beyond heuristic analyses of AR LMs. It offers a formal language and quantitative tools for investigating information flow, representation structure, and the mechanisms underlying the capabilities of these powerful models. While acknowledging limitations, particularly in estimating high-dimensional information quantities and fully formalizing the spectral learning connections (Section 7), this work lays theoretical groundwork. Future research directions include refining the spectral analogies, extending the analysis to multi-step dynamics, and exploring how these categorical insights might inform the design of more interpretable, controllable, or efficient language model architectures. Ultimately, this framework contributes to the development of a more principled theoretical understanding of deep generative models.

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2000.

[2] J. C. Baez, B. Fong, and B. S. Pollard. *A Compositional Framework for Markov Processes*. Journal of Mathematical Physics, 57(3):033301, 2016. arXiv:1508.06448 [math.PR].

[3] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, D. Hjelm. *Mutual Information Neural Estimation*. Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80:531-540, 2018.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems (NeurIPS), 33:1877-1901, 2020.

[5] K. Cho and B. Jacobs. *Disintegration and Bayesian Inversion via String Diagrams*. Mathematical Structures in Computer Science, 29(7):938–971, 2019. arXiv:1709.00321 [math.CT].

[6] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, et al. *A Mathematical Framework for Transformer Circuits*. Transformer Circuits Thread, Anthropic, 2021. https://transformer-circuits.pub/2021/framework/index.html

[7] T. Fritz. *A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics*. Advances in Mathematics, 370:107239, 2020. arXiv:1908.07021 [math.CT].

[8] J. Hewitt and C. D. Manning. *A Structural Probe for Finding Syntax in Word Representations*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pages 4129–4138, 2019.

[9] D. Hupkes, V. Dankers, M. Mul, E. Bruni. *Compositionality Decomposed: How do Neural Networks Generalise?*. Journal of Artificial Intelligence Research, 67:757-795, 2020.

[10] O. Kallenberg. *Foundations of Modern Probability*. Springer, 2nd edition, 2002.

[11] A. Kraskov, H. Stögbauer, P. Grassberger. *Estimating mutual information*. Physical Review E, 69(6): 066138, 2004.

[12] C. D. Manning, K. Clark, J. Hewitt, U. Khandelwal, O. Levy. *Emergent linguistic structure in pre-trained language models*. Proceedings of the National Academy of Sciences, 117(48):30046-30054, 2020.

[13] S. Nowozin, B. Cseke, R. Tomioka. *f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization*. Advances in Neural Information Processing Systems (NeurIPS), 29, 2016.

[14] C. Olsson, N. Elhage, N. Nanda, N. Joseph, D. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, et al. *In-context Learning and Induction Heads*. Transformer Circuits Thread, Anthropic, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

[15] P. Perrone. *Markov Categories and Entropy*. Entropy 25(1):110, 2023. arXiv:2212.11719v2 [cs.IT], 2023.

[16] P. Perrone. *Categorical Information Geometry*. In: Nielsen F., Barbaresco F. (eds) Geometric Science of Information. GSI 2023. Lecture Notes in Computer Science, vol 14072. Springer, 2023. arXiv:2306.05359 [math.CT].

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. *Language Models are Unsupervised Multitask Learners*. OpenAI Blog, 1(8), 2019.

[18] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, Y. Liu. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. arXiv preprint arXiv:2104.09864, 2021.

[19] N. Tishby, F. C. Pereira, W. Bialek. *The information bottleneck method*. Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing, pp. 368-377, 1999. arXiv:physics/0004057.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. *Attention is All You Need*. Advances in Neural Information Processing Systems (NeurIPS), 30, 2017.

[21] Q. Wang, S. R. Kulkarni, S. Verdú. *Divergence estimation for multidimensional densities via k-nearest-neighbor distances*. IEEE Transactions on Information Theory, 55(5): 2392-2405, 2009.

[22] A. Xu and M. Raginsky. *Information-theoretic analysis of generalization capability of learning algorithms*. Advances in Neural Information Processing Systems (NeurIPS), 30, 2017.

[23] Watanabe, Sumio. *Algebraic geometry and statistical learning theory*. Cambridge University Press, 25, 2009.

[24] H. Zhang, et al. *Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss*. Advances in Neural Information Processing Systems (NeurIPS), 34, 2021. arXiv:2106.04156.

[25] Z. Tan, Y. Zhang, J. Yang, Y. Yuan. *Contrastive Learning Is Spectral Clustering On Similarity Graph*. International Conference on Learning Representations (ICLR), 2024. arXiv:2303.15103.